基于Chunk-LDAvis的核心技术主题识别方法研究*

■ 刘自强[1,2] 许海云[1,3] 岳丽欣[4] 方曙[1]

[1] 中国科学院成都文献情报中心 成都 610041

[2] 中国科学院大学经济与管理学院图书情报与档案管理系 北京 100190

[3] 中国科学技术信息研究所 北京 100038 [4] 中国人民大学信息资源管理学院 北京 100872

摘要：［目的/意义］基于大量专利文献数据的核心技术主题识别有助于识别某技术领域的关键技术、分析关键技术的发展方向，是进行技术创新的基础情报工作，对于研究人员、企业乃至国家层面都具有一定的意义。［方法/过程］提出基于Chunk-LDAvis的核心技术主题识别方法，首先基于经典LDA模型进行主题识别，然后利用名词组块对初始LDA主题识别结果进行标注，构建Chunk-LDA主题识别结果，提高其可解读性；然后基于社会网络分析方法构建主题网络，识别核心技术主题；基于R语言的LDAvis工具包绘制可交互的Chunk-LDAvis核心技术主题关联分析图谱，发现核心技术主题的隐含联系，辅助进行核心技术主题识别。［结果/结论］通过对纳米农业领域进行实证研究，验证了本文提出方法的准确性和可行性。

关键词：Chunk-LDAvis 专利分析 主题识别 核心技术主题 交互可视化

分类号：G251.2

DOI:10.13266/j.issn.0252-3116.2019.09.008

随着新一轮的世界科技革命和产业变革的快速演进，技术创新持续涌现并促进新产品、新需求和新业态的产生，成为社会经济持续前进发展的关键驱动力，影响经济格局和产业形态的调整，成为驱动发展和提高国家竞争力的关键所在。目前，世界各国十分重视科技创新，纷纷加大在高新技术领域的投入，以期在新一轮的科技革命中抢占先机。我国近年来一直强调创新驱动发展战略，认为科技创新是提高社会生产力和综合国力的战略支撑，必须摆在国家发展全局的核心位置。

大数据时代，专利、论文等科技文献数量呈几何级数飞速增长[1]，支撑科技决策、科技创新的全局性、前瞻性、战略性的科技战略情报服务工作尤为重要。世界知识产权组织指出90%以上的科技信息是通过专利信息反映出来的，专利文献已经成为分析技术发展态势的重要、可靠的数据来源。尽管全球专利产出量呈飞速增长态势，但有关学者通过对欧洲国家20世纪50年代后专利价值的评估发现专利文献的价值分布并不均衡，约5%－10%的专利文献的价值占据了专利文献总价值的一半[2-3]，如何从大量的专利文献中准确、有效地捕捉到核心技术并预测其发展趋势成为目前科技情报工作中亟需解决的问题。

因此，国内外众多学者进行基于专利文献数据的技术识别与预测研究[4]，并取得了众多研究成果，例如：基于专利引文分析、专利主题词分析和可视化分析的核心技术识别与预测分析方法[5]，为各国的科技创新提供了一定的帮助，但是随着情报需求的不断深化，相应核心技术识别与预测方法有待进一步发展。

在目前研究的基础上，本文提出一种基于Chunk-LDAvis的核心技术主题识别方法，使之能够应对不断深化的科技创新情报需求，以期为科研人员、企业和国家等不同层面的科技情报工作提供一定的参考借鉴。

* 本文系国家自然科学基金项目“基于科学—技术主题关联分析的创新演化路径识别方法研究”（项目编号：71704170）和中国科学院成都文献情报中心青年人才创新项目（项目编号：Y7Z0581002）研究成果之一。

作者简介：刘自强（ORCID：0000－0003－1814－8655），博士研究生，E－mail：liuziqiang@mail.las.ac.cn；许海云（ORCID：0000－0002－7453－3331），副研究员，博士，硕士生导师；岳丽欣（ORCID：0000－0002－7268－7871），博士研究生；方曙（ORCID：0000－0002－4584－7574），研究员，博士生导师。

收稿日期：2018－07－10 修回日期：2018－11－04 本文起止页码：73－84 本文责任编辑：刘远颖

1　相关研究

1.1　基于引用特征的核心技术主题识别

随着世界科技革命和产业变革的快速演进,基于科技文献的核心技术识别研究广泛受到各国学者、企业和政府的高度重视。其中,国内外学者们就如何利用科技文献数据高效、准确地识别出核心技术、热点技术及其发展趋势展开了大量的研究工作。概括起来,主要可以分为两个方向,一是通过分析专利文献的同被引、耦合和直接引用等引用特征;二是通过分析专利文献的题名、摘要等文本内容特征进行核心技术主题识别。

其中,基于专利引用特征的核心技术主题识别方法较早受到学者的关注,例如:O. Kwon 等[6]通过构建专利引文耦合网络和共引网络,综合分析专利分布情况从而识别核心技术,并且通过 3 个领域的实证研究验证了该方法的有效性。C. Choi 等[7]提出一种基于主路径分析算法的核心技术识别方法,具体过程是:首先构建专利引文网络,然后利用主路径发现算法从中提取专利技术发展的主路径,最后通过分析技术演化脉络来识别技术领域的关键技术及其发展趋势。C. W. Hsu 等[8]利用专利聚类方法建立了生物制氢领域相关技术之间的相互引用矩阵,绘制了技术发展图谱并识别出其代表性技术领域。张欣等[9]结合专利的被引次数和专利的年龄对原始的 PageRank 算法进行改进,并将其应用到 OLED 领域中来识别核心专利。亢川博等[10]通过相互引证关系构建专利文献的引证网络,然后基于个体价值与网络价值指标进行核心专利主题识别。

基于引文特征的核心技术识别方法能够较为有效地识别核心技术,但由于引文分析存在引文时滞性(即一篇文献从发表到被引用需要一定的时间,而施引文献从完成到发表又需要一段时间),很多学者质疑基于引用特征识别核心技术的时效性、准确性,并且尝试深入专利文献内容进行挖掘,基于文本内容(专利题名、摘要等)特征进行共现、聚类分析,以期识别出更加具有可解读性、准确性的核心技术主题。

1.2　基于内容特征的核心技术主题识别

随着自然语言处理技术(文本聚类、LDA 主题模型[11]和社区识别[12-13]等)的发展,基于专利题名、摘要等内容特征的核心技术主题识别方法也逐渐受到学者的重视。

例如,Y. G. Lee 等[14]提出了一种应用于选择核心战略研究领域的“技术集群分析”方法,并将该方法应用于纳米技术领域的国家研发项目,具体思路是:关键词抽取、专利文档聚类、利用关键词在专利文档聚类中的层次分布关系分析核心技术,并利用该方法预测了韩国纳米技术领域的三大核心技术集群。栾春娟等[15]以德温特专利库为数据源,抽取“德温特指南代码”(Derwent Manual Code, DMC)并绘制共现网络进行可视化分析,从而识别核心技术领域,最后以航空航天领域为例进行了实证研究。范宇等[16]提出了适用于专利信息聚类的主题模型和聚类算法,将潜在狄利克雷分配(LDA)主题模型和 OPTICS 算法相结合进行核心专利主题分析。李佳佳等[17]利用社会网络分析方法对中国、美国和欧洲等不同国家的专利分类号共现网络图进行对比分析,识别出中国、美国和欧洲的核心专利领域。伊惠芳等[18]结合 LDA 模型和战略坐标图方法进行专利技术主题分析,识别出其中的核心技术主题及其结构特征,对于客观合理地追踪技术前沿、提高研发效率具有重要意义。

虽然,基于专利文献文本内容(关键词、分类号等)进行共现、聚类分析相比基于引用特征的方法具有一定的优势(不存在引文时滞性);但同样存在一定的不足,如关键词之间欠缺语义关系,无法反映词与词之间的关联关系,而且不能有效揭示技术主题之间的关联关系。

1.3　LDA 模型的改进与应用

LDA 模型最早是由 D. M. Blei 等于 2003 年提出,可以基于统计概率层面表达词间语义层次关系[19]。2006 年,D. M. Blei 等又提出了动态主题模型,让 LDA 模型可以处理具有时间戳记的文档数据集,实现动态主题识别与追踪[20]。但 D. M. Blei 等提出的经典 LDA 模型存在一定不足,例如,LDA 识别结果中每个主题是一组单词,不便于解读;主题识别之后,主题—主题、主题—词语之间关联如何衡量。

针对这两点不足,有关学者进行了改进研究,取得了众多研究成果,如 TNG(Topical N-Grams)模型[21]、PhraseLDA 模型[22-23]和 LDAvis 模型[24],其中 TNG、PhraseLDA 模型采用短语表示主题,具有更好的语义表达性;LDAvis 模型能够基于多维尺度算法将主题识别结果映射到二维空间中,进而揭示主题—主题、主题—词语之间的关联。近年来,图情领域的学者利用 LDA 模型及其改进算法进行科研主题识别,将其应用于基于文本数据的科技情报分析工作中,范云满等[25]利用 TNG 模型进行了新兴主题的探测研究。张琴

等[26]利用 PhraseLDA 模型进行了主题短语挖掘方法的研究，结果表明该方法在多种数据集中挖掘出的主题短语质量较高。

在上述分析基础上，本研究借鉴 TNG、PhraseLDA 模型，采用名词组块(Chunk)表示主题(相较于短语 Phrase，名词组块 Chunk 语义信息含量更高)，然后利用 LDAvis 模型揭示主题—主题、主题—词语之间的关联，进而构建 Chunk-LDAvis 模型，并将其应用于核心技术主题识别研究中。基于 Chunk-LDAvis 进行核心技术主题识别，一方面可以将每个核心技术主题表示为一组名词组块，提高可解读性，另一方面可以揭示核心技术主题、主题词之间的相互关联。

综上所述，针对目前利用专利文献数据进行核心技术识别研究中的不足，本文提出基于 Chunk-LDAvis 模型的核心技术主题识别方法，主要创新之处在于通过名词组块标注进行语义增强的 LDA 主题识别，并基于 Web 前端技术研究探索交互式可视化技术进行主题关联分析，从而提升核心技术主题识别分析的准确性和可读性，并通过对整个流程的实证研究验证该方法的有效性。

2 基于 Chunk-LDAvis 的核心技术主题识别框架

通过对相关研究的总结归纳，核心技术主题识别研究存在两个相互联系的改进、提升方向：①增强技术主题的语义信息，提高内容特征信息量；②识别技术主题之间内容维度的关联，并利用关联关系识别核心技术主题。前者是基础，即通过主题模型、语义分析等方法能够更加有效地(相较于关键词、引文链接)归纳、概括专利文本的内容特征；后者是深化，即在技术主题语义表征的基础上，可以增加核心技术主题之间语义维度的关联，而不是简单的共现关联，进而依据语义维度的关联关系识别核心技术主题。基于以上分析，本文提出基于 Chunk-LDAvis 的核心技术主题识别框架，主要包括数据收集与处理、语义增强的主题识别、核心技术主题判定和关联可视化分析等 4 个系统流程，主要思路如图 1 所示：

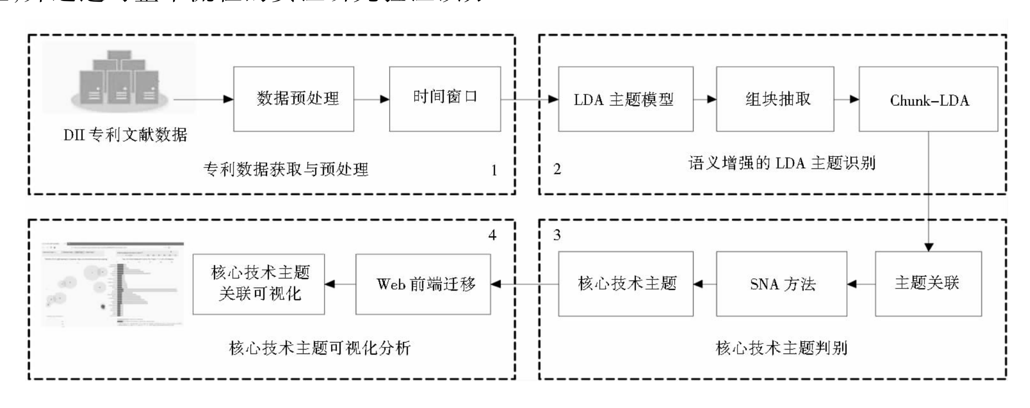


图 1 基于 Chunk-LDAvis 的核心技术主题识别的方法思路

第一步，数据收集与处理。根据目标确定数据库，构建检索式获取相应技术领域的专利文献。然后进行数据处理，包括专利文献格式转换，因为研究目的为核心专利主题分析，所以需要进行题名、摘要和时间等关键信息提取，并保存至本地以备后续研究使用。

第二步，语义增强的 LDA 主题识别。首先基于经典 LDA 模型进行主题识别，然后利用词性标注、句法分析和语法分析进行主语名词组块(Subject Noun Trunk，即代表主语的名词短语)、宾语名词组块(Object Noun Trunk，即代表宾语的名词短语)抽取，并以之为基础对初始 LDA 主题识别结果进行标注，构建 Chunk-LDA 主题识别结果，进而增强 LDA 主题识别结果的语义功能，提高其可解读性。

第三步，核心技术主题判别。划分时间窗口，构建主题网络，基于社会网络分析方法识别核心专利主题。

第四步，基于 Chunk-LDAvis 的核心技术主题可视化分析。利用 Web 前端技术，绘制可交互的 Chunk-LDAvis 核心技术主题关联分析图谱，然后搭建 Web 数据库进行在线测试，从核心技术主题识别结果的语义提升和可解读性等两个方面进行改进，进而有效识别、分析核心技术主题。

下面对主要步骤进行详细介绍。

2.1 语义增强的 LDA 主题识别

(1)初始 LDA 主题识别。近年来学界提出了众多主题模型，比如潜在语义索引[27](Latent Semantic Analysis, LSA)、概率性潜在语义索引[28](Probabilistic La-

tent Semantic Analysis，pLSA）和 LDA 模型等。与 LSA 和 pLSA 模型相比，LDA 模型不仅能预测训练集文档的主题分布，而且能够有效预测非训练集中的文档和词的主题分布，因此，LDA 模型逐渐成为分析大规模非结构化文档集的最有效工具之一。

具体来讲，LDA 是一种三层（词、主题和文档）贝叶斯概率模型（见图 2），LDA 模型假设文档是由若干隐性主题组成，而主题是由词表中的所有词汇组成。LDA 主题模型的联合分布概率如公式（1）所示：

$$P(\theta,z,w)=P(\theta|w)\prod_{n=1}^{N}P(z_n|\theta)P(w_n|z_n,\beta)$$

公式（1）

其中，M 为文档数目，K 为主题数目，N 表示第 m 个文档的单词数目，θ 为参数 α 的 Dirichlet 分布采样，z 表示主题，w 表示主题词，φ 为参数为 β 的 Dirichlet 分布采样。

LDA 模型生成过程可以概括为以下步骤：

1）从参数为 β 的 Dirichlet 分布中为每个主题采样主题—单词分布 φ_k，即有 $k\varphi_k\sim$ Dir（β），$k\in[1, K]$。

2）从参数为 α 的 Dirichlet 分布中为每个文档采样文档—主题分布 θ_m，即有 $\theta_m\sim$ Dir（α），$m\in[1, M]$。

对文档 m 中第 n（$n\in[1, N_m]$）个词：

3）从参数为 θ_m 的多项式分布中采样 1 个主题 $z_{m,n}$，即有 $z_{m,n}\sim Mult(\theta_m)$。

4）从参数为 $\varphi_{zm,n}$ 的多项式分布中采样 1 个具体单词 $w_{m,n}$，即有 $w_{m,n}\sim Mult(\varphi_{zm,n})$。

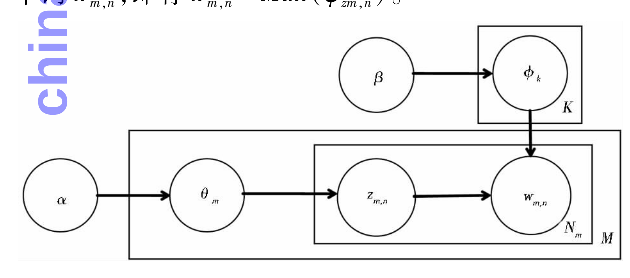


图 2　LDA 模型

本研究具体使用 R 语言下的 topic models 工具包（https://cran.r-project.org/web/packages/topicmodels）进行 LDA 主题识别。在 R 语言中，主要有两个工具包提供了 LDA 模型：LDA 和 topic models 工具包。前者提供了基于 Gibbs 采样的经典 LDA、MMSB（the Mixed-meMbership Stochastic Block Model）、RTM（Relational Topic Model）和基于 VEM（Variati-onal Expectation-max-imization）的 sLDA（supervised LDA）、RTM。后者提供 LDA_VEM、LDA_Gibbs、CTM_VEM（correlated topics model）3 种模型。

利用 LDA 对文献数据进行建模关键是要推断出超参数 α 和 β，即要计算出每个文档—主题分布 θ_m 和主题—单词分布 $\varphi_{zm,n}$ 隐式参数。目前对于 LDA 模型中参数估计的方法有最大后验估计 MAP（Maximum a Posteriori）、变分贝叶斯估计 VB（Variational Bayes）、变分贝叶斯推断 CVB（Variational Bayesian Inference）和吉布斯采样 GS（Gibbs Sampling）等方法，本研究选用 R 语言下的 topic models 工具包的 LDA_Gibbs 模型对 LDA 模型参数进行估计。

（2）语义组块标注。在初始 LDA 主题识别处理之后，对于各个主题的支持文档，基于 Python 语言，利用词性标注、句法分析和语法分析抽取某一主题下的支持文档中代表主、宾语的名词组块。具体可以分为 TAG、CHUNK 和 ROLE 3 个步骤。

首先 TAG，根据各个词在句子中的作用，对其进行词性标注，主要包括动词（VB）、名词（NN）、代词（PR＋DT）、形容词（JJ）、副词（RB）、介词（IN）、连词（CC）和感叹词（UH）等。

CHUNK，即进行组块（chunk）标注，组块标签分配给属于在一起的单词组（即短语），比如名词短语（NP，例如 the red coat）和动词短语（VP，例如 is doing），具体如表 1 所示：

表 1　组块标签及其含义

组块标签	含义	成分	例子
NP	名词短语	DT＋RB＋JJ＋NN ＋ PR	the strange bird
PP	介词短语	TO＋IN	in between
VP	动词短语	RB＋MD＋VB	was looking
ADVP	副词短语	RB	also
ADJP	形容词短语	CC＋RB＋JJ	warm and cosy
SBAR	从属连词	IN	whether or not
INTJ	感叹词	UH	hello

ROLE，语义角色标签描述了不同组块之间的关系，阐明了组块在句子中的作用。句子中最常见的角色是 SBJ（主语名词短语）和 OBJ（宾语名词短语）。句子的主语是做某事或做某事的人、事物、地点或想法。句子的宾语是受动作影响的人/物，具体如表 2 所示：

表 2　组块语义角色标签及其含义

语义角色标签	含义	成分	例子
SBJ	主语名词短语	NP	the boy sat on the Chair
OBJ	宾语名词短语	NP＋SBAR	the boy sat on the Chair

直观解释上述过程，以“Phrase-*LDAvis model is helpful to detect the core technology topic.*”这一句子为例进行语义组块抽取测试，结果见图 3，可以标注出每个

单词的词性、组块以及标注组块的语义角色,最终得到该句子中代表主语成分的名词组块 Phrase-LDA model 和代表宾语成分的名词组块 the core technology topic,如图 3 所示:

```
IPython console
Console 1/A
In [1]: runfile('C:/Users/lenovo/Desktop/组块抽取示例/测试.py', wdir='C:/Users/lenovo/Desktop/组块抽取示例')
         WORD   TAG    CHUNK   ROLE   ID     PNP    LEMMA

 Phrase-LDAvis   NN     NP      SBJ    1      -      phrase-ldavis
         model   NN     NP ^    SBJ    1      -      model
            is   VBZ    VP      -      1      -      be
       helpful   JJ     ADJP    -      -      -      helpful
            to   TO     VP      -      2      -      to
        detect   VB     VP ^    -      2      -      detect
           the   DT     NP      OBJ    2      -      the
          core   JJ     NP ^    OBJ    2      -      core
    technology   NN     NP ^    OBJ    2      -      technology
         topic   NN     NP ^    OBJ    2      -      topic
             .   .      -       -      -      -      .
Chunk('Phrase-LDAvis model/NP-SBJ-1') => [Chunk('Phrase-LDAvis model/NP-SBJ-1')]
Chunk('the core technology topic/NP-OBJ-2') => [Chunk('the core technology topic/NP-OBJ-2')]
```

图 3　语义组块抽取测试结果

(3)Chunk-LDA 标注。在第(2)步语义组块抽取结果的基础上,对第(1)步 LDA 主题识别初始结果主题词(单一主题词)进行组块标注,从而实现对 LDA 主题识别结果的组块标注,比如,以"Phrase-*LDAvis model is helpful to detect the core technology topic*"中的主题词 technology 进行组块标注:technology → the core technology topic,可以得到语义增强的 Chunk-LDA,从而提高主题识别结果的可读性(语义功能)。

由于某一主题词可能对应若干个主、宾语名词组块,该步骤中关键问题在于如何确定主题词对应的组块。本文采取的解决方法是,首先根据主题词对应的主题来确定相应的生成文档(主题词—主题—主题文档),然后抽取出这些对应文档的语义组块,并按照频率排序,再以主题词为线索词选择出对应的频次最高的组块,从而完成 Chunk-LDA 构建。

2.2　基于 SNA 的核心技术主题识别

专利文献中蕴含的技术主题之间存在或明显或隐含的联系,而这种联系可以揭示某一技术主题的重要程度和核心价值,比如技术主题 T 与其他若干主题联系越多,表明主题 T 的核心性越高。

目前 LDA 主题识别方法,虽然可以识别出大量文本中的主题,但是无法分析哪些主题属于核心主题。因此,本研究中尝试基于社会网络分析(Social Network Analysis,SNA)方法对 LDA 主题识别结果做进一步处理:即在 LDA 主题识别结果的基础上,构建 LDA 主题社会网络图,通过中心性指标判断核心技术主题,中心性计算方法如公式(2)所示:

$$C_i(T) = \frac{1}{\lambda}\sum_{j=1}^{n} A_{ij} C_j \qquad \text{公式 (2)}$$

其中,$C_i(T)$ 为主题 Ti 的中心性,该中心性计算公式为 Bonacich's Centrality,即特征向量中心性[29](eigenvector centrality);A_{ij} 为网络的邻接矩阵,λ 为常数,C_j 为 C_i 节点的邻接点。

例如,基于 LDA 模型识别出的若干技术主题集合标记为 T = {topic 1,topic 2,topic 3,…,topic n},然后基于 R 语言的 igraph 工具包进行主题网络 G 构建,并计算各个节点的中心性值 Ci(T),并将其中心性值的大小通过主题节点大小进行表示,如图 4 所示:

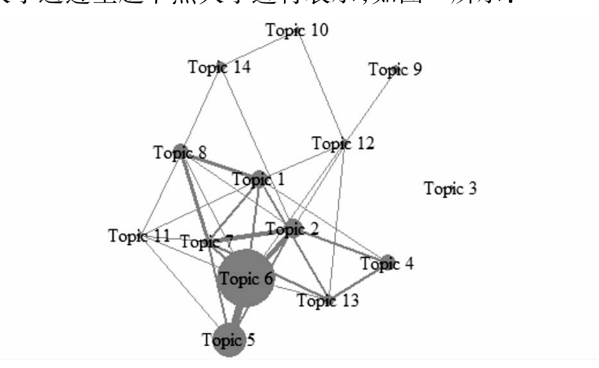
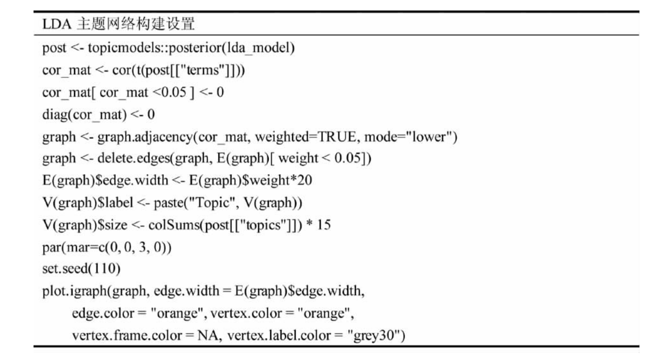


图 4　基于 SNA 的核心技术主题网络示意

具体处理工具是基于 R 语言的 igraph 工具包进行主题网络 G 构建,可视化布局设置代码如图 5 所示:


LDA 主题网络构建设置

```
post <- topicmodels::posterior(lda_model)
cor_mat <- cor(t(post[["terms"]]))
cor_mat[ cor_mat <0.05 ] <- 0
diag(cor_mat) <- 0
graph <- graph.adjacency(cor_mat, weighted=TRUE, mode="lower")
graph <- delete.edges(graph, E(graph)[ weight < 0.05])
E(graph)$edge.width <- E(graph)$weight*20
V(graph)$label <- paste("Topic", V(graph))
V(graph)$size <- colSums(post[["topics"]]) * 15
par(mar=c(0, 0, 3, 0))
set.seed(110)
plot.igraph(graph, edge.width = E(graph)$edge.width,
    edge.color = "orange", vertex.color = "orange",
    vertex.frame.color = NA, vertex.label.color = "grey30")
```

图 5　LDA 主题网络构建设置

2.3　基于 Chunk-LDAvis 的核心技术主题可视化分析

科技情报分析应该以用户为导向,但是目前核心技术主题识别研究结果主要以本地静态图谱展示,难以多层次、细粒度地分析核心技术主题内容,往往只能观看到情报分析人员提供的内容。随着信息技术的发展,如交互式可视化技术可以在一定程度上弥补这一不足,即可以通过交互式的可视化结果多层次地展示科技情报结果,满足用户的个性化需求。

而且,上一步中虽然识别出了核心技术主题,但是核心技术主题与其他主题的相关关系和具体内容(主题的下位词)无法明确得到,因此,需要做进一步分析。本研究基于多维尺度分析[30](Multidimensional Scaling,MDS),利用主题间的欧氏距离,去构建低维空间,

使得 LDA 主题在此空间的距离和在高维空间中的 LDA 主题之间的相似性尽可能地保持一致,主题之间距离的远近表示主题的相关性,可以用这种方式来进一步分析核心技术主题。

基于上述分析,本研究探索利用 R 语言的 LDAvis 工具包(https://github.com/cpsievert/LDAvis)来绘制交互式的核心技术主题可视化图谱,其基本可视化布局及含义如图 6 所示:

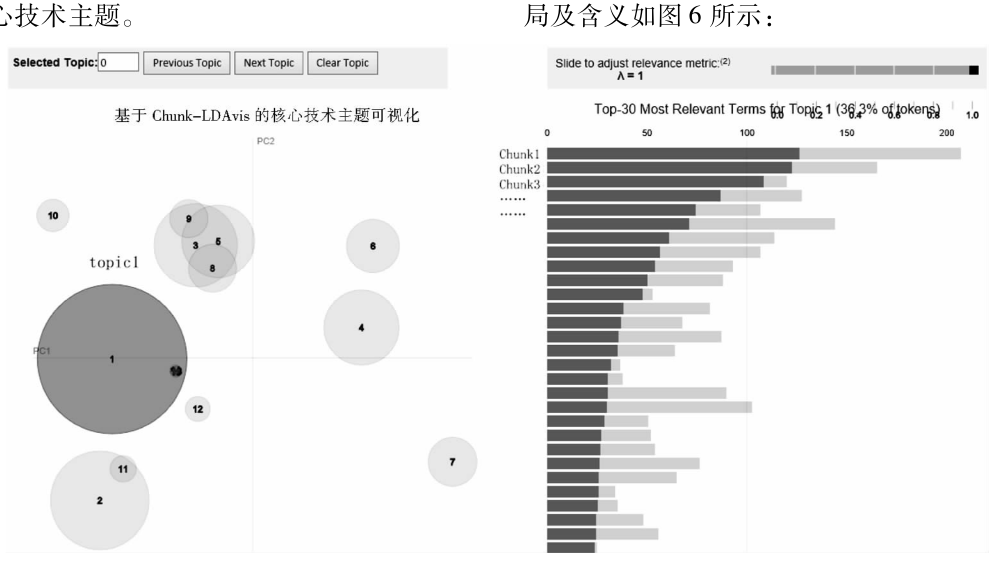


图 6 LDA 主题可视化示意

图 6 主要可以分为左右两个部分,左边是基于 MDS 算法将 LDA 主题可视化展示在二维空间中,圆点表示主题(圆点中的数字为 LDA 主题识别结果序号),圆点的大小由主题对应文档多少决定;右边为主题对应的词项,按照生成概率大小排序。该图利用 Web 前端工具生成,具有良好的交互式可视化效果。以 topic1 为例,点击 topic1 圆点,右边会交互式地展示 topic1 下位词项。点击右边的某一词或者组块可以相应显示对应的主题 topic。基于上述处理步骤结果,可以进行较为充分、直观的核心技术主题分析。

此外,还可以通过调节参数 λ(0 ≤λ≤ 1)来控制主题—词语关联度 relevance(term w | topic t),即可以控制显示某一主题的不同的下位词项。参数 λ 计算方法如公式(3)[24]所示:

$$r(w,k|\lambda) = \lambda\log(\varphi_{kw}) + (1-\lambda)\log(\frac{\varphi_{kw}}{p_w}) \quad 公式(3)$$

其中,w 表示主题词,$w \in \{1, 2, 3, \cdots, V\}$; k 表示主题,$k \in \{1, 2, 3, \cdots, K\}$; φ_{kw} 表示 Gibbs 采样参数; p_w 表示主题词 w 的分布概率。

$\lambda = 0$ 时,显示主题下特有的、相对独立的下位词项,即这些词项往往只出现在该主题; $\lambda = 1$ 时,显示分布概率更高的下位词项,但是这些高分布概率的词项往往不单独属于该主题,也会同时属于其他主题。

2.4 特点与优势

本研究构建的基于 Chunk-LDAvis 的核心技术主题识别框架和基于引用特征、基于文本内容特征的核心技术主题识别方法相比,具有以下特点与优势:

从结果准确性层面上来看,是对基于经典 LDA 模型的核心技术主题识别方法的改进(通过主题相关文档数量判断核心技术主题,认为某技术主题相关文档数量越多越可能是核心技术主题),增加了主题关联视角的核心技术主题判别维度。

从结果内容层面上来看,每个核心技术主题是由一组名词组块构成,相较于一组单词或专利号等语义表达能力更强,便于用户进行解读。

从结果呈现方式层面来看,相较于静态的核心技术主题知识图谱,以动态、交互式的可视化图谱形式呈现,对用户更加友好,便于进行情报分析。

3 实证研究

3.1 数据源

本文以德温特创新索引(Derwent Innovations Index, DII)数据库所收录的 2010 年 1 月 1 日 -2017 年 12 月 31 日纳米农业领域的专利数据为数据源。DII 数据库是基于 Web 的专利信息数据库,收录了来自全球 40 多个专利机构(涵盖 100 多个国家)的 1 000 多万条基本发明专利,2 000 多万条专利信息,有利于在一个技术主题下进行全球专利研发状况和技术攻关信息的研究,因此,以 DII 数据库收录的纳米农业专利数据作为识别纳米农业领域核心技术主题的数据源是可行有效的。

在 DII 数据库中,使用检索式 Keyword = “Nano agriculture *”进行检索,检索时间跨度为 2010 年 1 月 1 日 -2017 年 12 月 31 日,得到检索结果 4 937 项。各年度专利数量如图 7 所示:

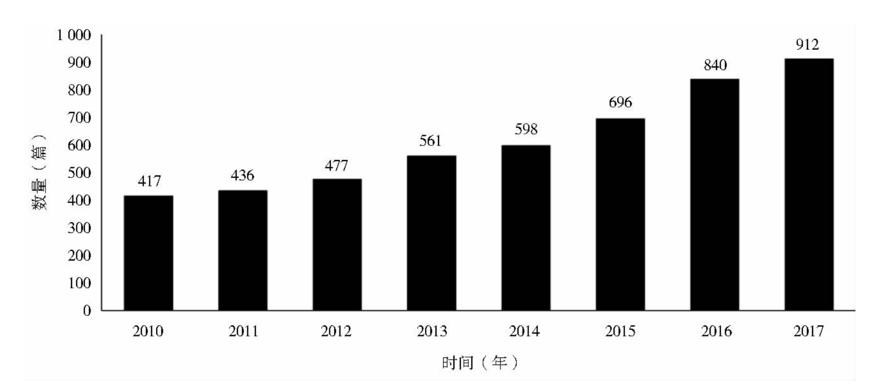


图 7　专利数量年度划分

3.2　语义增强的 LDA 主题识别

进行 LDA 主题识别首先应该做的是预估输入文档集合中共有多少个主题(K),目前研究中,学者主要利用困惑度(perplexity)和对数似然值(log likelihood)的变化进行估计。前者随着主题数量的增加递减,后者随着主题数量的增加递增,一般两者变化趋于平缓时的主题数即可作为估计的主题数量。其中,本研究使用对数似然值(log likelihood)进行最优主题数的确定。

目前研究中,最优主题数确定之前需要对数据集中包含的主题数目有一定的先验估计。本研究估计所下载的专利数据集中的主题数目为 100 个以内。因此,进行迭代实验以确定最优主题数,K 从 1 - 100,步进为 25,每个主题数运行 1 000 次迭代,得到每个 K 和对应的对数似然值,如图 8 所示:

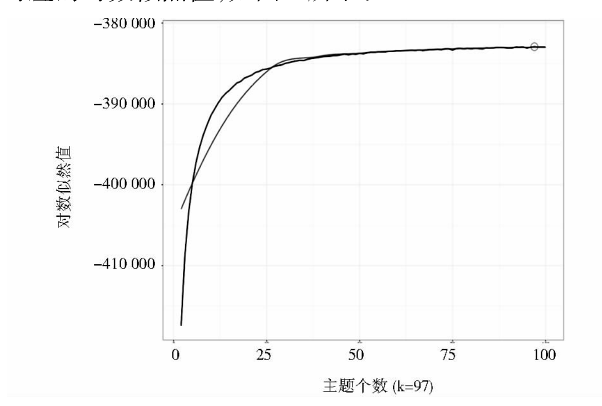


图 8　基于对数似然值的 LDA 主题个数确定

从图 8 中可以看出,当主题数取 90 时 LDA 模型的对数似然值趋于稳定,并且于 97 得到最大值。因此,本次实验选择主题数为 K = 97。在完成 LDA 主题个数确定的基础上,利用 R 语言下的 topic models 工具包进行 LDA 主题识别,然后将主题识别结果保存至本地以为其添加组块标注。初始 LDA 主题识别结果得到之后,按照前文所述方法利用 Python 进行组块抽取,组块抽取部分结果见图 9。

然后,基于组块抽取结果,对初始 LDA 主题识别结果进行组块标注,得到语义增强的 LDA 主题(Chunk -LDA 主题),并以 Chunk -LDA 主题—文档矩阵的形式保存至本地,部分结果见图 10。

3.3　基于 SNA 的核心技术主题识别

在上一步数据处理结果(语义增强的 LDA 主题识别)的基础上,基于社会网络分析(Social Network Analysis,SNA)方法对 LDA 主题识别结果做进一步处理:在 LDA 主题识别结果的基础上,构建 LDA 主题社会网络图,通过中心性指标判断核心技术主题。对识别出的 97 个主题进行社会网络分析,构建 LDA 主题可视化网络,结果如图 11 所示,其中节点的大小由中心性决定。

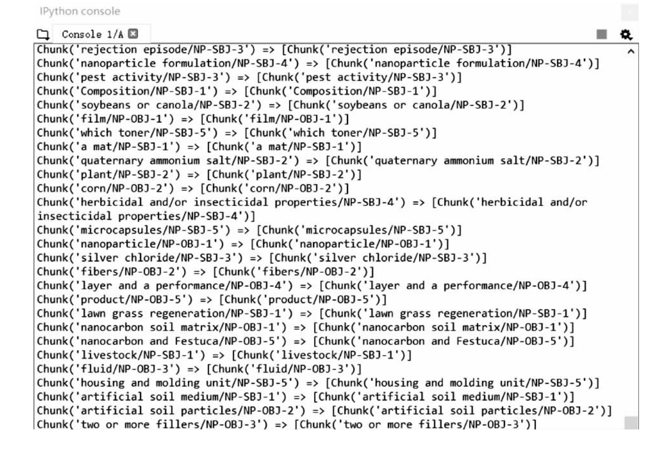


图 9　组块抽取结果部分

通过计算、排序,得到核心技术主题排序及其 Ci(T)值,具体见表 4。结合可视化结果,通过分析中心性可以较为直观地发现纳米农业领域的核心主题,如 Topic1、Topic40、Topic60 等主题位于纳米农业领域主题网络的核心位置,进而可以判断其为核心技术主题。

基于 SNA 方法虽然可以识别出纳米农业领域的核心技术主题,但是其解读与分析还存在一定的难度,难以满足实际情报分析需求,因此,本研究对其结果做进一步处理,基于 Web 前端技术将结果进行交互式可视化处理,增强结果的可读性和分析维度。最后,基于核心技术主题可视化图谱,分析纳米农业领域的核心

技术主题。

A	B	C	D	E	F
Document_Topic	comprises nanotitanium, strain cqma421, emulsifier water, agent solvent, polymer carrier, wrapped phospholipid, foodstuff beverage, stabilizer solvent, additive comprises,	selenium enriched, herbicidal composition, preparation method, powder comprises, filling wood, fiber material, electromagnetic radiation, metarhizium anisopliae, method producing,	selenium content, lipid nanoparticles, agent sterilizing, substrate comprising, nanoparticles stabilizing, comprises nanoparticulate, specific selenium, beauveria bassiana, solution prepared,	promoting agent, powdery mildew, nanosilver antimicrobial, rich organic, nanocomposite antibacterial, dichlorophenoxyacetic acid, comprises nanostructured, comprises nanotitanium, antimicrobial applications,	silver antibacterial, essential nanoliposome, phospholipid bilayer, connected lamp, agent additive, containing polymerassociated, nanoparticle containing, antibacterial activity, agent including, surface
1	0. 057377049	0. 073770492	0. 040983607	0. 040983607	0. 040983607
2	0. 053846154	0. 069230769	0. 053846154	0. 053846154	0. 053846154
3	0. 040983607	0. 040983607	0. 040983607	0. 040983607	0. 040983607
4	0. 056451613	0. 040322581	0. 040322581	0. 072580645	0. 040322581
5	0. 0703125	0. 0390625	0. 0703125	0. 0390625	0. 0703125
6	0. 043859649	0. 043859649	0. 043859649	0. 096491228	0. 043859649
7	0. 040322581	0. 040322581	0. 040322581	0. 072580645	0. 040322581
8	0. 044642857	0. 044642857	0. 044642857	0. 044642857	0. 044642857
9	0. 03968254	0. 071428571	0. 03968254	0. 03968254	0. 03968254
10	0. 040983607	0. 057377049	0. 057377049	0. 040983607	0. 040983607
11	0. 0390625	0. 0390625	0. 0546875	0. 0390625	0. 0546875
12	0. 048076923	0. 048076923	0. 048076923	0. 048076923	0. 048076923
13	0. 058333333	0. 041666667	0. 041666667	0. 041666667	0. 041666667
14	0. 038461538	0. 038461538	0. 053846154	0. 053846154	0. 038461538
15	0. 041666667	0. 058333333	0. 041666667	0. 041666667	0. 041666667
16	0. 049019608	0. 049019608	0. 049019608	0. 049019608	0. 049019608
17	0. 053846154	0. 053846154	0. 038461538	0. 053846154	0. 053846154
18	0. 049019608	0. 049019608	0. 049019608	0. 049019608	0. 049019608
19	0. 038461538	0. 084615385	0. 038461538	0. 053846154	0. 038461538
20	0. 048076923	0. 067307692	0. 048076923	0. 048076923	0. 048076923
21	0. 0390625	0. 0859375	0. 0703125	0. 0390625	0. 0546875
22	0. 056451613	0. 040322581	0. 040322581	0. 040322581	0. 056451613
23	0. 059322034	0. 059322034	0. 076271186	0. 059322034	0. 042372881

图 10　Chunk-LDA 主题识别结果(部分)

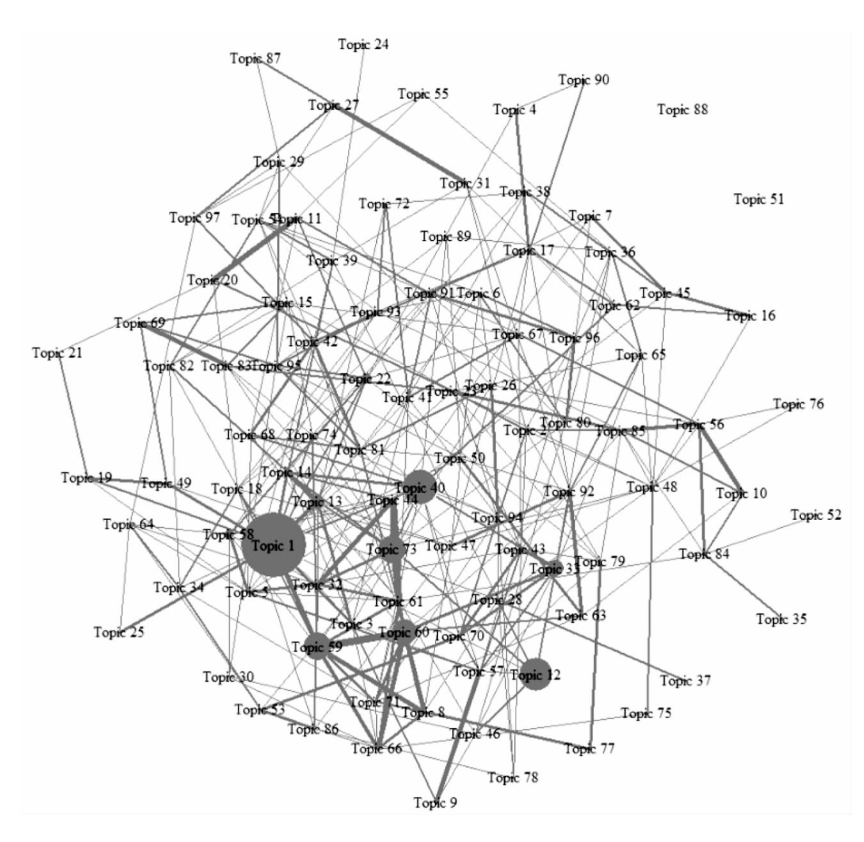


图 11　纳米农业领域核心技术主题网络

表 4　核心主题排序及其 Ci(T)值(部分)

排序	主题 T	Ci(T)	排序	主题 T	Ci(T)	排序	主题 T	Ci(T)
1	Topic 1	1. 00	7	Topic 33	0. 76	13	Topic 9	0. 27
2	Topic 40	0. 87	8	Topic 6	0. 56	14	Topic 14	0. 25
3	Topic 59	0. 86	9	Topic 5	0. 34	15	Topic 22	0. 24
4	Topic 73	0. 82	10	Topic 15	0. 31	16	Topic 50	0. 21
5	Topic 12	0. 79	11	Topic 13	0. 29	17	Topic 92	0. 16
6	Topic 60	0. 78	12	Topic 8	0. 29	18	Topic 44	0. 16

3.4　基于 Chunk-LDAvis 的核心技术主题可视化分析

在上一步基于 SNA 的核心技术主题识别结果的基础上,选取排序前 15 的核心技术主题,利用 LDAvis 工具包来绘制交互式的纳米农业领域核心技术主题可视化图谱,图 12 为纳米农业领域核心技术主题可视化静态结果,动态、可交互的可视化结果已经上传到自建网站,可以在线访问(https://www. informationscience. top/core technology topic/)。网页中左边代表主题编号的圆点可以点击,鼠标停留在圆点上会显示构成该主题的 Top－30 名词组块;右边的名词组块也可以点击,鼠标停留在名词组块上会显示该名词组块所在的主题。

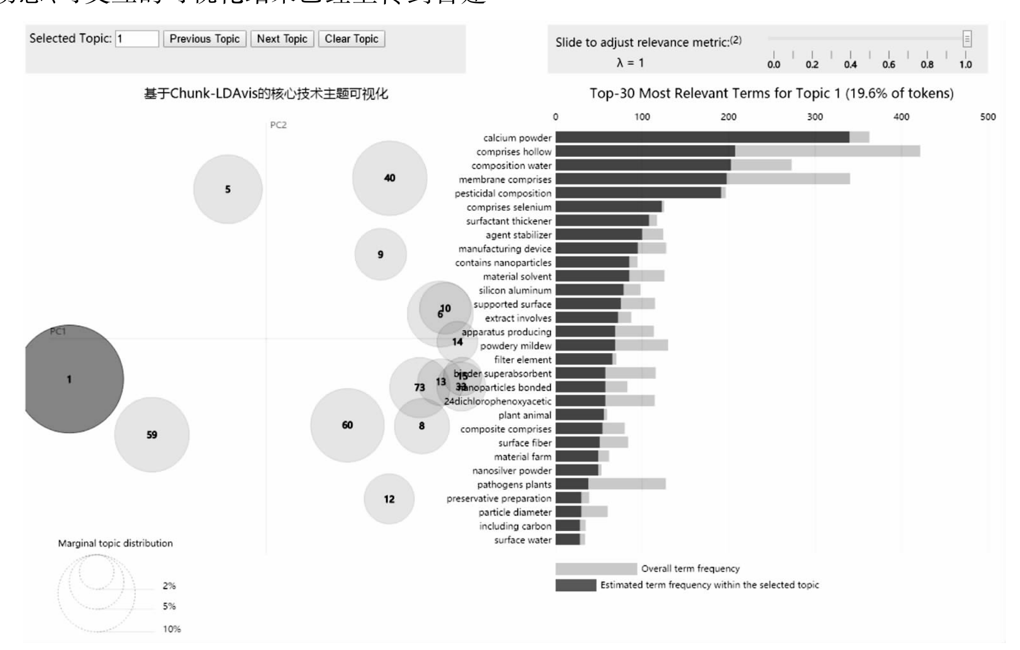


图 12　纳米农业领域核心技术主题可视化

在图 12 中,可以发现 Topic 1、Topic 40、Topic 59、Topic 60、Topic 73、Topic 12 和 Topic 33 等 7 个主题核心主题的位置与 LDA 主题网络图中的位置基本一致,但是图 12 中节点的大小正比于主题概率,因此和 LDA 主题网络图中主题节点的大小(正比于主题中心性)有所差异。

在上述结果的基础上,综合进行纳米农业领域的核心技术主题分析,选取纳米农业领域的 top3 的核心技术主题进行具体分析:

(1)Topic1－纳米农药。结合 Topic 1 下位短语对其进行分析可知,纳米技术在农药中的应用主要有 3 种形式:①利用纳米加工技术使农药原药纳米化,制成纳米分散体、纳米乳剂、纳米颗粒或纳米微球,增加农药制剂的比表面积,提高农药的油溶性或水混溶性,改善农药在水中的分散性和稳定性,促进吸收。此类纳米农药有噻虫啉农药纳米颗粒、唑菌胺酯农药、氟氯氰菊酯纳米乳剂组合物以及一些苯并唑、苯基化合物等纳米微粒。②利用纳米载体负载农药,提高环境敏感农药的稳定性,改善药物在作物表面的粘附性和渗透性,减少流失。③将一些金属或无机材料制入农药,增强农药的杀菌和光催化作用,促进农药分解,降低农药残留。如新型光触媒杀虫剂、纳米二氧化钛复合农药等。另外,一些新型纳米农药和防虫害缓释剂还可以增加植物的害虫抗性或真菌抗性,抑制微生物的生长和增值,保障植物的健壮性,并具有良好的除草效率和环保性。

(2)Topic 40－农业装置与器械。通过分析该主题的具体内容,可以发现纳米技术在农业装置与器械方面的应用主要集中于以下 3 个方面:①灌溉系统、净水系统与养殖系统。纳米技术在灌溉、净水与养殖等运用和处理水的多系统中的应用主要体现于使用纳米管进行排水和净水的纳米净化曝气器、纳米气泡发生装置等,用于保温的纳米碳布以及用于承重和容纳的纳米支撑盘和纳米盘槽。②温室大棚装置。纳米技术在温室大棚装置中的应用集中于使用纳米碳管收集太阳能热力,使用纳米碳布和纳米玻璃进行保温,使用纳米涂层进行发电和杀菌,使用纳米发电玻璃和纳米电网进行照片等。③自走式联合收割机、播种机、施肥机。

纳米技术在此类农业器械中的应用主要集中于发挥支撑、受力和保护作用的纳米机架和纳米挡板、发挥运输作用的纳米输送系统、发挥采集作用的纳米管和纳米袋、发挥去污功能的纳米纤维等。

（3）Topic 59－农业环境改良。分析该主题的具体内容可知，纳米材料因其巨大的比表面积以及可修饰的多种官能团使其容易与环境中的有机化合物和重金属粒子等污染物结合，在农业环境改良方面发挥着越来越重要的作用。如目前研究中侧重利用氧化锌/硅藻土纳米复合材料进行污水处理；纳米二氧化钛用于生物吸附剂，可以具备优异的吸附容量、较高的重金属选择性和较高的降解去除有机污染物、病原菌和微生物的能力；如何利用氧化石墨烯和氧化铁磁性纳米颗粒制成磁性纳米杀菌剂也是该主题的重要内容。利用纳米技术将银纳米化，纳米银杀菌具有光谱抗菌、强效杀菌、渗透性强、抗菌持久等特点，在农业环境改良方面尤其是污水处理和抗菌杀毒方面应用广泛。在污水处理方面，该主题主要研究二氧化钛纳米复合水凝胶软化反应器进行污水软化，使用二氧化钛光催化剂去除藻类，使用纳米银复合材料进行有机毒物的分解降解及其杀菌应用等。在抗菌杀毒方面，主要研究使用银水性聚氨酯来抗菌、杀菌和除臭，使用银纳米颗粒稳定悬浮液、将银离子吸附到二氧化硅纳米粒子上来作为杀生物剂或通过混合羟基吡啶酮化合物和银化合物制备杀生物组合物可有效抑制和消除水中微生物。

3.5 核心技术主题识别结果的检验

将实证结果与具体纳米农业领域专利分析实践工作的结果进行对比，以检验本研究提出的核心技术主题方法的可行性和有效性。在具体实践工作中（纳米农业领域专利态势调研分析工作，原始数据相同），使用科睿唯安旗下的 TI（Thomson Innovation，TI）绘制了纳米农业领域专利地图（见图 13）。专利地图中山峰的海拔高度代表特定主题文献的密度大小，并显示不同主题之间的相对关系，可以用于核心技术主题分析。

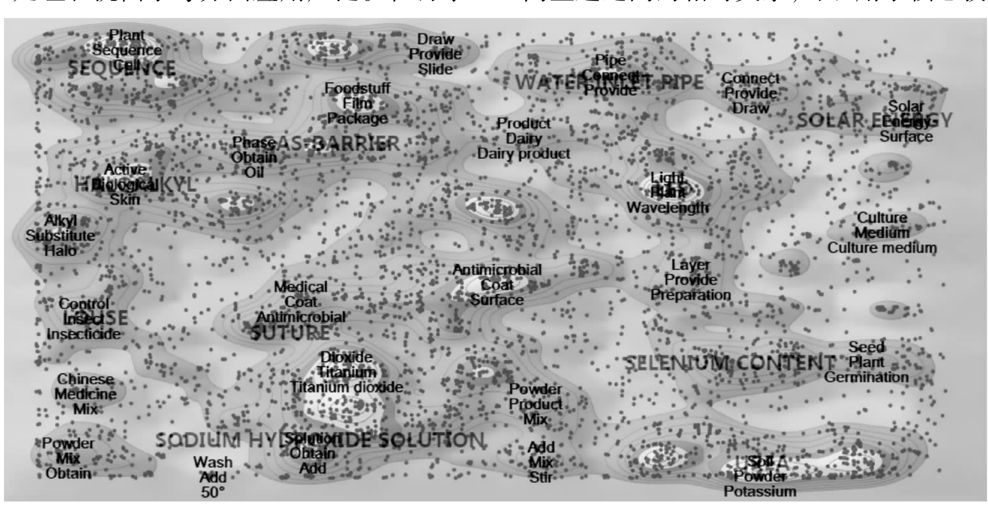


图 13 纳米农业领域专利地图

分析图 13，可以发现纳米农业领域的核心技术主题主要包括农药、肥料、农业装置与器械、农产品加工、农业种植与栽培、农业环境改良和动植物遗传育种与纳米检测等 7 个主题。通过与本研究得到的纳米农业领域核心技术主题识别结果对比检验，可以发现本研究识别出的核心技术主题 Topic 1－纳米农药、Topic 40－农业装置与器械和 Topic 59－农业环境改良识别结果与专利地图结果 1、3 和 6 相符合，在一定程度上可以检验本文提出方法的可行性与有效性。

3.6 讨论

与基于经典 LDA 模型的核心技术主题识别方法相比，本文提出的方法，一方面通过语义组块标注改进了经典 LDA 结果中单一主题词的语义信息不足问题；另一方面，相较于单纯依靠主题分布概率高低来判断核心主题，提出基于社会网络和多维尺度分析识别主题之间的关联关系及其可视化的方法。与目前基于关键词和分类号共现的核心专利主题分析方法相比，本文中提出的基于 Chunk-LDAvis 的核心技术主题识别方法，更加具有针对性、可读性（不是单一的关键词或者分类号，基本知识单元为表示主语或者宾语的名词组块），而且能够交互式可视化展示、分析某技术领域核心技术主题，提高了识别结果的可读性。

但是本方法也存在一定的局限，如关于解决核心技术主题语义信息不足这一问题，本文通过构建 Chunk-LDA 主题来解决，由于其通过半人工的方法得到，在分析效率上存在一定的不足。因此，需要探索更

加有效的机器学习方法,实现 Chunk-LDA 主题自动化构建。此外,基于语义 TRIZ 的专利主题表征方法也可以解决目前核心技术主题识别研究中语义信息不足的问题,即将基本专利知识单元表示为 SAO(主谓宾)结构,再通过划分不同维度,可以实现宏观、中观和微观的多层次核心技术主题分析。

4 结语

本文在调研总结核心技术主题识别方法的基础上,提出基于 Chunk-LDAvis 的核心技术主题识别方法,可以用来分析某专利领域的核心技术主题。创新之处主要有两点:一是提出一种新的基于语义组块标注的 LDA 主题分析方法,二是利用 Web 前端技术实现了对核心技术主题的隐含关系的可视化分析。最后以纳米农业领域为例,选取 2010 年至 2017 年间共 4 937 篇专利文献作为数据源,利用本文提出的核心技术主题识别方法进行了实证研究,证明本文提出的方法是可行、有效的。但是,本文提出的核心技术主题识别方法还存在两点主要不足:①Chunk-LDA 主题通过半人工的方法得到,当待分析专利数据量过大时会存在分析效率低的不足;②无法充分分析核心技术主题发展趋势。因此,接下来的工作是进行 Chunk-LDA 主题自动化构建以及核心技术主题演化路径识别研究,实现对核心技术主题的动态追踪。

参考文献:

[1] 王效岳,白如江. 海量网络学术文献自动分类技术研究[M]. 北京:人民出版社,2015:40-42.

[2] SCHANKERMAN M, PAKES A. Estimates of the value of patent rights in European countries during the post-1950 period[J]. Economic journal, 1986, 96(384):1052-1076.

[3] 许海云,岳增慧,雷炳旭,等. 基于专利技术功效主题词与专利引文共现的核心专利挖掘[J]. 图书情报工作,2014,58(4):59-64.

[4] 袁润,钱过. 识别核心专利的粗糙集理论模型[J]. 图书情报工作, 2015,59(2):123-130.

[5] 马永涛,张旭,傅俊英,等. 核心专利及其识别方法综述[J]. 情报杂志,2014,33(5):38-43,70.

[6] KWON O, SEO J, NOH K, et al. Categorizing influential patents using bibliometric analysis of patent citations network[J]. Information-an international interdisciplinary journal, 2007, 10(3): 313-326.

[7] CHOI C, PARK Y. Monitoring the organic structure of technology based on the patent development paths[J]. Technological forecasting and social change, 2009, 76(6):754-768.

[8] HSU C W, CHANG P L, HSIUNG CM, et. al. Charting the evolution of biohydrogen production technology through a patent analysis [J]. Biomass & bioenergy, 2015,76(5):1-10.

[9] 张欣,马瑞敏. 基于改进 PageRank 算法的核心专利发现研究[J]. 图书情报工作,2018,62(10):106-115.

[10] 亢川博,王伟,穆晓敏,等. 核心专利识别的综合价值模型[J]. 情报科学,2018,36(2):67-70.

[11] WANG Y, BAI H J, STANTON M, et al. PLDA: parallel latent Dirichlet allocation for large-scale applications[C]//International conference on algorithmic aspects in information and management. San Francisco: Springer-verlag, 2009:301-314.

[12] NEWMAN M E J, GIRVAN M. Finding and evaluating community structure in networks[J]. Physical review, 2004,69(2): 108-113.

[13] BLONDEL V D, GUILLAUME J L, LAMBIOTTE R, et al. Fast unfolding of communities in large networks [J]. Journal of statistical mechanics: theory and experiment,2008,30(2):155-168

[14] LEE Y G, SONG Y I. Selecting the key research areas in nano-technology field using technology cluster analysis: a case study based on National R&D Programs in South Korea[J]. Technovation, 2007, 27 (12):57-64.

[15] 栾春娟,曾国屏. 基于 SNA 核心技术领域测度研究[J]. 图书情报工作,2011,55(6):33-35.

[16] 范宇,符红光,文奕. 基于 LDA 模型的专利信息聚类技术[J]. 计算机应用,2013,33(S1):87-89,93.

[17] 李佳佳,马铁驹. 基于专利数据的风能核心技术识别及趋势分析[J]. 科技管理研究,2017(12):129-136.

[18] 伊惠芳,吴红,马永新,等. 基于 LDA 和战略坐标的专利技术主题分析——以石墨烯领域为例[J]. 情报杂志,2018,37(5):97-102.

[19] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation [J]. Journal of machine learning research,2003(3): 993-1022.

[20] BLEI D M, LAFFERTY J. Dynamic topic models[C]//Proceedings of the 23rd international conference on machine learning. NewYork:ACM,2006:113-120.

[21] WANG X, MCCALLUM A, WEI X. Topical N-Grams: phrase and topic discovery, with an application to information retrieval[C]// IEEE international conference on data mining. Omaha: IEEE Computer Society, 2007:697-702.

[22] ELKISHKY A, SONG Y, VOSS C R, et al. Scalable topical phrase mining from text corpora[J]. Proceedings of the VLDB endowment, 2014, 8(3):305-316.

[23] LI B, WANG B, ZHOU R, et al. CITPM: A cluster-based iterative topical phrase mining framework[C]// International conference on database systems for advanced applications. Dallas: Springer International Publishing, 2016:197-213.

[24] SIEVERT C, SHIRLEY K. LDAvis: a method for visualizing and interpreting topics[C]// Proceedings of the workshop on interactive language learning, visualization, and interfaces. Baltimore: Association for Computational Linguistics, 2014:63-70.

[25] 范云满,马建霞. 基于 LDA 与新兴主题特征分析的新兴主题探测研究[J]. 情报学报,2014,33(7):698 – 711.
[26] 张琴,张智雄. 基于 PhraseLDA 模型的主题短语挖掘方法研究[J]. 图书情报工作, 2017, 61(8): 120 – 125.
[27] LANDAUER T K, DUMAIS S T. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge[J]. Psychological review, 1997, 104(2):211 – 240.
[28] SHEN C, LI T, DING C H Q. Integrating clustering and multi-document summarization by bi-mixture probabilistic latent semantic analysis (PLSA) with sentence bases[C]// AAAI conference on artificial intelligence. San Francisco: AAAI Press, 2011:914 – 920.
[29] BONACICH P B. Factoring and weighting approaches to status scores and clique identification[J]. Journal of mathematical sociology,1972, 2 (1):113 – 120.
[30] STURROCK K, ROCHA J. A multidimensional scaling stress evaluation table[J]. Field methods, 2016, 12(1):49 – 60.

作者贡献说明:
刘自强:设计论文整体研究框架,撰写论文;
许海云:提出论文研究思路,指导论文修改;
岳丽欣:撰写论文结果分析部分;
方曙:指导论文修改。

Research on Core Technology Topic Identification Based on Chunk-LDAvis

Liu Ziqiang[1,2] Xu Haiyun[1,3] Yue Lixin[4] Fang Shu[1]
[1] Chengdu Library of Chinese Academy of Sciences, Chengdu 610041
[2] Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190
[3] Institute of Scientific and Technical Information of China (ISTIC), Beijing 100038
[4] School of Information Resource Management, Renmin University of China, Beijing 100872

Abstract: [**Purpose/significance**] Core technology topic identification based on a large number of patent documents is helpful to detect key technologies in a technical field and to analyze the direction of the development of key technologies. It is the basic information work for technological innovation and has certain significance for researchers, enterprises and even the national level. [**Method/process**] This paper proposes a core technology topic identification method based on Chunk-LDAvis. Firstly, it is based on the classic LDA model to identify the topics. Then, the noun chunk is used to mark the results of the initial LDA topic identification, and the result of the Chunk-LDA topic recognition is constructed to improve its interpretability. Then based on the social network analysis method, the topic network is constructed to identify the core technical topics; based on the LDAvis toolkit, the interactive Chunk-LDAvis core technology topic association analysis map is plotted, and the hidden links of the core technical topics are found, and the core technology topic detection is assisted. [**Result/conclusion**] Through the empirical study on the field of nanoscale agriculture, the accuracy and feasibility of the proposed method are verified.

Keywords: Chunk-LDAvis patent analysis topic recognition core technology topics interactive visualization